VDW VEREINIGUNG DEUTSCHER WISSENSCHAFTLER

Policy Paper on the Asilomar Principles on Artificial Intelligence

Research Team

Technology Assessment of Digitisation

Authors (Members of the VDW Research Team)

Prof. Dr. Ulrich Bartosch

Prof. Dr. Stefan Bauberger SJ

Tile von Damm

Dr. Rainer Engels

Prof. Dr. Malte Rehbein

Frank Schmiedchen (Leader)

Prof. Dr. Heinz Stapf-Finé

Angelika Sülzen

Imprint:

Publisher: Federation of German Scientists e.V. (VDW) © 2018 Office of the Federation of German Scientists, Marienstraße 19/20, 10117 Berlin

June 2018

VDW VEREINIGUNG DEUTSCHER WISSENSCHAFTLER

Table of contents

1. Introduction

The Federation of German Scientists e.V. (VDW), through its research team "Technology Assessment of Digitisation", presents its first statement on the ethical, political and legal questions of research, development and application of Artificial Intelligence (A.I.), emphasizing general or strong artificial intelligence. It is based on the Asilomar Principles on Artificial Intelligence adopted on January 6, 2017, which well reflects the current state of the discussion started by the A.I. developer community on the above questions.[1] A very large group of globally relevant actors involved in research and development of artificial intelligence identify with these principles. These include many of the key players in Western industrialized countries, some of whom hold leading positions.

The research, development and application of artificial intelligence can pose a fundamental threat to mankind and its peaceful coexistence. It is in the public interest to recognize such developments at an early stage and to enable reasonable actions to avert danger. It is part of the principle of hazard prevention to act preventively before developments that lead to a hazard have reached an irreversible momentum that no longer permits effective countermeasures. Progress in the field of A.I. creates such a fundamental threat. This has been recognized by leading specialists in this field of research and has led to the formulation of the Asilomar principles, among other things. This is highly acknowledged and appreciated.

Our review of the Asilomar principles has shown that they are not consistent with the logic of successful hazard prevention. If the containment of the A.I. were only to follow the recommendations of the Asilomar principles, in our view, an essential risk would be accepted, which is equally important to other threats to humanity and for which the necessity and logic of unconditional preventive security is already accepted.

The nuclear military strategy worldwide still follows this logic, for example. With a few exceptions, it is still recognized today that it is necessary to prevent the use of (also tactical) nuclear weapons,

[1] Future of Life Institute (2017).

because an uncontrollable escalation to the self-destruction of mankind threatens to lead to a war of extinction without victors. Since mutual trust could not and cannot be assumed as a guarantee against hostile action, it is necessary to reliably combine mutual controllability with the simultaneous potential for mutual destruction and to implement a concept of "common security".[2]

Global climate policy also follows this logic. Since 1992, the international community has had to limit a further increase in the earth's temperature. The dangerous situation has been clearly recognized and concerted action is needed to prevent a "self-immolation" of humanity.[3] Since the interests in the field of climate policy are far more diverse and ramified than in the case of the threat of war from nuclear weapons, it is far more difficult to reach a consensus for global community political action. The scientifically unequivocally diagnosed facts of anthropogenic climate change are often outplayed by conflicting economic, political and cultural interests. Nevertheless, the transformative agenda of the Paris Convention on Climate Change[4] and the Sustainable Development Goals (SDG)[5] reflects the logic of unconditional preventive hazard prevention for this threat to humanity.

In the field of biotechnology, too, irreversible interference with the natural genome of humans, animals and plants is increasingly endangering the whole of humanity. The progressive process of changing human nature and nature as human environment is widely accepted even if this causes uncontrollable further effects. Above all, scientific curiosity, optimism about progress, business interests and claims to power often overlap. The latest developments make it possible for organic hackers in "garage laboratories" to produce dangerous organisms according to "recipes" and remove the technological application from any state control. With Gene Drive, the "laws of evolution" were broken and a dynamic of unimaginable proportions was set in motion. Leading scientists in this field

[2] VDW-Member Egon Bahr. Bahr/Lutz (1992) und Independent Commission on Disarmament and Security Issues (1982).
[3] VDW-Member Hans Joachim Schellnhuber. Schellnhuber (2015).
[4] United Nations (2015a).
[5] United Nations (2015b).

demand: "Protect society from our inventions".[6] "It is a real threat to humanity. And the papers don't report it.[7]

What these fundamental threats have in common is the dynamic that leads to a loss of control over their further developments. Nuclear warfare, climate catastrophe, genetic evolutionary change all have the potential to extensively destroy human life on Earth. At the same time, this potential danger is hidden behind the perceived control sovereignty of the instigator of this development: human kind. It is therefore necessary to counter those dangers that are barely visible or not yet visible (at least to the majority of observers). Public attention is most sensitized to the implications of nuclear developments. The dangerous to catastrophic extent of man-made climate change is also a widespread concern. The discomfort with the interventions in the evolutionary processes is rather low. A new threat of a similar epochal magnitude that has only become realistic in the last three years is also largely unknown to the public. It is the subject of this VDW statement.

Artificial intelligence has opened Pandora's box as well. It has the potential to undermine the logic of control - including the dangers mentioned above. In the digital age, the tools of human thought may unfold an autonomous position that is powerfully directed against the impotent human being. It is high time for a reflected, active defense against such danger. This is the aim of the following text.

The Federation of German Scientists supports this warning. Its foundation dates back to the warning of the "Eighteen of Göttingen"[8] in 1957 against the dangers of nuclear armament.[9] The responsibility of science today means no less than it did then to make the unrecognized dangers visible to the public and to serve security with scientific expertise. With this in mind, the VDW “Research Team Technology Assessment of Digitisation" has started its work and presents its first results.

[6] Oye, et.al. (2014).
[7] VDW-Member Ernst-Ulrich von Weizsäcker. Weizsäcker/Wijkman (2017).
[8] Göttinger Achtzehn (1957).
[9] https://vdw-ev.de/ueber-uns/geschichte-und-ziele/.

2. Objectives, definitions and initial theses

The aim of the opinion is to contribute to the European and international discussion on the possible consequences of A.I. and commensurate countermeasures. As a VDW research group, we argue from the point of view of the protection of individual, economic, social and cultural human rights[10], and also consider the possible consequences of a general or strong A.I. for humanity as a whole. The EU's precautionary principle is the legal compass for consideration.[11]

The term intelligence is controversial. We therefore use a generally accepted basic understanding that describes intelligence as a not directly observable phenomenon that describes a person's cognitive abilities and knowledge available to a person at a given time.[12]

Artificial intelligence is an (empirical) branch of computer science and deals with methods that enable a computer to solve tasks that, when solved by humans, require intelligence. The term "artificial intelligence" was first used in 1956 by the US computer scientist John McCarthy.[13] As early as 1950 Alan Turing described the possibility of intelligence simulated by computers.[14] Definitional features of A.I. are the built-in ability to learn and to deal with uncertainty, inaccuracy/fuzziness and probabilities right from the start.[15]

Weak A.I. is understood as such forms of A.I. in which machines simulate intelligent behavior with the aid of mathematics and computer science that are limited to a specific area of application and are capable of learning. By contrast, general or strong A.I. implies a general learning ability, including the ability to develop autonomously. A superintelligence is defined as a strong A.I. that is superior to the human brain at least in many areas. It arises hypothetically when a strong A.I. improves and expands

[10] United Nations (1948) und United Nations (1966).
[11] Commission of the European Communities (2000).
[12] Cognition is understood here as 'situated cognition', which means not only internal calculation processes in the brain but also and above all the reciprocal real-time interaction of a physically composed system with its environment in a certain way (Walter, 2014).
[13] BITKOM (2017).
[14] EFI (2018), p. 69.
[15] BITKOM (2017) describes a taxonomy of the automation of decision-making and develops a 5-step model for this purpose.

itself through recursion.[16] Possible development directions for the realisation of a superintelligence are algorithms (also in machines, robots, etc.), transhumanism (e.g. genetic improvements of humans or the fusion of both in cyborgs) as well as the use of artificial, neural networks.

Artificial intelligence has a varied history. By now, it has reached a stage of development in which it can drastically change all areas of life. The starting point of our work is the concern that the probable dangers of a general or strong artificial intelligence will not be adequately discussed in time and thus adequate security measures will become impossible. On the one hand, this is due to the fact that the major breakthroughs in weak A.I. have only taken place in the past three years and on the other hand, as with other risk technologies (e.g. nuclear power, climate change, genetic engineering), the possible dangers of the still non-existent strong A.I. are abstract, invisible and unknown. Data and algorithms only unfold their effect when they enter and manifest themselves in the human reality of life.

The following initial hypotheses form the basis of the opinion:

1. Research, development and application of A.I. are developing exponentially.[17] Alongside the USA and China, Germany is a leader in basic research.[18]

2. Even applications of weak A.I. change social and communication behavior[19] and our everyday culture (e.g. social scoring) can threaten social systems (e.g. electoral influences).

3. The creation of strong A.I. is probable in view of the exponential growth rate of A.I. research, even if the time of realization is difficult to predict.[20]

[16] The strong A.I. succeeds in adapting and improving its own program code in such a way that it can make further improvements in the next stage or its next generation, which would have been impossible in the previous versions. These recursive learning cycles are continued by the strong A.I. until human intelligence is surpassed and becomes superintelligence. A prerequisite for this is that the strong A.I. "understands" its purpose and the design of its own program code.
[17] BITKOM (2017); EFI (2018).
[18] EFI (2018), p. 68ff.
[19] Eberle (2015), Henk (2014). However, the influence of A.I. is difficult to measure independently of other aspects of digitization.
[20] Bostrom/Müller (2013): These range from 2022 (median optimistic year), over 2040 (median realistic year) to 2075 (median pessimistic year). Similar estimates can be made on the basis of the Moore´sche Act of 1965,

4. The transition from weak to strong A.I. is a continuous process, which takes place mainly where the necessary resources (data, finances, power) are available. The results of research and development (R&D) in this area are not necessarily published due to economic and political interests, which makes socially legitimate control impossible or at least considerably more difficult.

5. A strong A.I. can become dangerous for individuals and for humanity through its autonomous action:

a. An A.I. makes decisions intended by man for the achievement of goals or self-preservation, which harm people as a side effect (collateral), which can also lead to the degree of complete subjugation or annihilation of humanity, without the A.I. having "consciously wanted" it.[21]
b. An A.I. is destructively intended (e.g. as a deadly autonomous weapon system) and increases its efficiency in a way that endangers more people/humanity.
c. An A.I. develops unintended competencies and pursues self-imposed goals that endanger individuals or humanity.

6. From an indefinite point in time, people can no longer follow the self-improvement process of a strong A.I. (not even as an "interconnected" collective), since their learning progress is always slower than that of the A.I., making control/correction impossible.

7. From the initial hypotheses 3 to 6 arises the necessity for the implementation of the precautionary principle.[22] With a sufficient time gap before the creation of a strong A.I., all necessary (above all

extrapolating the explosive progress of the last three years. According to Raymond Kurzweil, the computing power of the human brain is around 10,000 teraflops. This computing power has already been significantly exceeded by large computer systems. In addition, they can be directly networked with each other. But there are also voices that claim that the development of strong A.I. "is not yet feasible in the foreseeable future" (EFI (2018), p. 69).

[21] In further development of an example by Stephan Hawking: Responsible people plan and build a hydropower plant. They have carried out extensive environmental and social impact assessments. Mammals and bird nests were evacuated and people were generously compensated for the necessary resettlement before the reservoir was flooded. But nobody has taken care of the 50,000 ant colonies that have now drowned at the bottom of the reservoir: 2,500,000,000,000 ants murdered are the result "which no one wanted". Hawking says that we must act now if we want to avoid one day possibly being in place of the ants.

[22] According to Agenda 21 (Chapter 35, Paragraph 3) adopted at the UN Conference on Environment and Development in Rio (1992), the precautionary principle can be defined as a compelling call to action, possible risks/hazards which, with an (exactly) predictable probability of occurrence, can lead to (exactly) unpredictable

normative) steps must be successfully completed that prevent a strong A.I. from endangering individuals and humanity as a whole in any form, at any time, under any circumstances (anticipatory governance). To achieve this, targeted decisions and appropriate measures must be taken without delay.

8. Man and machine begin to separate in three ways:

a. Technology increasingly acts independently and thus takes over the functions and tasks of people;
b. neurological-cognitive components are incorporated into humans (so-called "enhancements");[23]
c. Coupling of human brains and A.I. systems.

We can speak of a progressive dissolution of boundaries with strong ethical implications. The issues related to transhumanism are addressed only to some extent in this opinion.

3. Critical reflection of the Asilomar principles

3.1 Introduction

The Future of Life Institute (FLI), founded in Boston (USA), has been dealing with the possible existential dangers of further technological development for mankind since March 2014. The work on risk reduction of A.I. is explicitly the focus of the Institute's work.[24] In January 2017, the FLI organized the "Beneficial AI" conference in Asilomar (on the Californian coast) with almost 1000 participants, including over 100 of the world's leading A.I. researchers and entrepreneurs, in order to discuss the

negative consequences for people (especially life and health) and the environment, now or in the future, to take all targeted protective measures preventively. This also ties in with Hans Jonas' ethics of responsibility.

[23] TAB (2016) and Vorwinkel (2017).

[24] The founders and strong supporters of the institute are Stephan Hawking, Elon Musk, Max Tegmark (MIT), Jaan Tallinn (Skype inventor), Stuart J. Russell (computer science), George Church (biology), Saul Perlmutter and Frank Wilczek (physics) as well as Alan Alda and Morgan Freeman (actors). In January 2015, Elon Musk funded a USD 10 million research program on K.I. for the FLI, which focuses on security issues and the development of "useful" A.I. The funds were used to fund 37 research projects.

effects of A.I. The "Asilomar AI Principles" are a result of this conference. The program committee was composed of people who have an immediate or indirect professional interest in A.I. [25] The 23 principles adopted and signed by numerous scientists represent a proposal for a voluntary commitment for research, development and application of A.I. They can be considered being a reaction to the accelerated technological development in this area, which the organizers of the FLI rightly call "major change [...] across every segment of society".[26] Within the first six weeks, the principles had already been signed by more than 1000 scientists directly involved in research and development of A.I. and by almost 2000 other individuals.

They have already shown initial effects:

- The University of Montreal began an open process in 2017 to develop the Montreal Responsible AI Principles, which should be completed in 2018.[27]
- Some 60 scientists from 30 countries have called for a boycott of the South Korean KAIST Institute, because of its cooperation Hanwha, a South Korean armaments group.[28]

The motto of the above-mentioned "Beneficial AI" conference has already defined the approach and is also reflected in the short preamble to the principles: "Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead."[29] With this purely positivist and utilitarian view of the use of A.I. technologies, the big and perhaps decisive open question remains (among others) : How to deal with developments that are not beneficial for everyone or not at all and, above all, how to deal with the threats posed by these developments.

[25] Future of Life Institute (2017).
[26] ibid.
[27] Université de Montréal (2018).
[28] On the establishment of the „Research Centers for the Convergence of National Defense and Artificial Intelligence“. Centre on Impact of AI and Robotics (2018).
[29] Future of Life Institute (2017).

The assessments in this opinion are based on the postulate of compliance with the United Nations Convention on Human Rights as an absolute, albeit not sufficient, minimum condition and apply the EU legal understanding of the precautionary principle as a set of values and norms to be observed. Furthermore, the considerations refer to Hans Jonas' ethics of responsibility as a philosophical reference, which is also the basis of the precautionary principle.[30]

According to Jonas' approach of the "heuristics of fear", every human decision must first be based on the potential consequences for the future that this decision could entail. Jonas' motive to "preserve the integrity of his world [of human kind] and his being against the encroachments of his power"[31] and his imperative "Act in a way that the effects of your action are compatible with the permanence of real human life on earth"[32] are also a helpful benchmark for the evaluation of A.I. Since from this point of view we can only derive how we do not want to live, positive-normative proposals must also be made on how the handling of A.I. should be structured.

So what is the world in which we want to live in the future? The Asilomar principles do not answer this question. They assume a generally valid and widely accepted consensus of a technic optimistic conception of the future, which on the one hand remains undetermined and runs the risk due to a lack of socio-economic analysis that only a few will determine it, but on the other hand, with the further pursuit of the technology path, accepts A.I. as an inevitable fate. Taken together, the two already pose social threats to democracy, the rule of law and human rights.

- Who determines what is "good" when the technology is virtually all-encompassing and affects everyone - not just those who use a particular A.I.-based product?
- Do we have a consensus on the hazards/risks we are prepared to accept in order to benefit from A.I.?
- How can such a consensus be achieved on a global scale?
- Will this consensus once established continue to exist in the future when things that result from the consensus can no longer be reversed?

[30] Jonas (1979).
[31] Ibid, p..9.
[32] Ibid, p. 35

- If a set of rules to control strong A.I. have been established, who can guarantee that the strong A.I. itself will not override this set of rules and set its own standards (including ethical standards) - and that it will do so at a speed that would make effective human (counter-) reactions impossible?

The Asilomar principles address a number of ethical issues related to A.I. and describe morally derived best practices related to A.I. research and development (R&D), while allowing a wide scope for interpretation. The principles use numerous undefined legal terms which would have to be defined if they are to be developed as a manageable instrument. This raises the question, who has the right to define.

Many chosen formulations seem to assume that the scientists entrusted with R&D by A.I. work together in a domination-free, cooperative manner, or that this is possible, provided that the researchers are willing to do so. This already raises the questions:

- Is this a realistic starting point?
- To what extent must it be considered from the outset that R&D also takes place for A.I. in contexts in which external objectives (e.g. entrepreneurial gain, national security) have at least a significant impact or essentially determine the research agenda?
- Is it realistic that the use of A.I. can be controlled solely by voluntary agreements between researchers without formal participation of existing institutional structures and processes of the democratically constituted political space?

The Asilomar principles are divided into the following sections: Questions of Research, Ethics and Values, and Longer-Term Problems.

3.2 Research questions

An essential starting point of the principles is the demand that R&D on A.I. is focused so that only "useful" A.I. is created. The question of who defines how, if and when an A.I. is useful remains

unanswered. The connection between "directed" and "beneficial", constructed in principle No. 1, is illogical, since both are different categories.[33] Something can be undirected or uncontrolled, but still useful, just as there are many things that arise in a controlled way, but are useless or harmful. The aim of A.I. research should therefore be both to control the "intelligence" and to ensure that it acts always and everywhere ecologically and socially sustainable.

Building on this, the fundamental problems of utilitarianism have to be discussed, since the purpose does not justify all means - or justifies them only in situations of extreme danger - particularly if there is not (yet) a common basic understanding of the purpose, but the purpose is perhaps only dominated by a few (e.g. certainly deadly autonomous weapon systems are very "useful" from the point of view of their producers).

So, who determines for which people, groups or institutions the A.I. must at least be useful in order to be considered useful overall? Thus A.I. can lead to economic efficiency gains, but the question whether we want these efficiency gains socio-culturally can not only be answered by the sum of consumer decisions. We have reached a point of technological development at which it must be questioned for each individual case whether an increase in convenience does not demand too high a price in terms of the loss of human abilities and skills.[34] If we decide that the “net” efficiency gains are wanted, a second step is to decide on their distribution by society. These questions must be addressed primarily by the cultural, social and economic disciplines.[35]

In general, a legitimate research policy is needed that clearly defines what ethically responsible innovation in the field of A.I. means in detail.

Principle No. 2 deals with the need for accompanying research. Like other technical developments, A.I. systems have an impact on social development processes and have a fundamental influence on them. For this reason, it is necessary to include issues of irreversibility and risk assessment in research to

[33] “The goal of AI research should be to create not undirected intelligence, but beneficial intelligence”.
[34] A well-known example for decades is the socially recognized "self-torture" in the gym, because our bodies are not challenged enough in everyday life.
[35] There is no evidence to date that the use of K.I. will lead to a reduction in social inequality. In contrast, however, there is initial evidence that it tends to promote exactly the opposite. This generally applies to the digitization of economic processes and thus also to A.I.

better understand all relevant ethical and societal challenges. This must also be done for weak A.I. and especially before the development of a strong A.I. and the relevant research must be adequately funded. The EU has currently put forward initial proposals which will have to be considered carefully.[36]

Principle No. 3 speaks of a "constructive" and "healthy" exchange between A.I. researchers and political decision-makers.[37] In contrast, the principle of democracy provides that the legislature and the executive are not on an equal level with groups of industry representatives or scientists. Rather, objectives must be democratically legitimized and violations of operational rules for the implementation of these objectives must be able to be sanctioned by the state and, in the case of cross-border aspects, also multilaterally.[38]

Principles No. 4 and 5 assume that it is possible to create a culture of cooperation, trust and transparency among A.I. researchers and developers.[39] In this general formulation, this is only realistic on a personal level. Only in a few (particularly public) institutions researchers are paid to work together in a spirit of trust. Otherwise, this behavior will only be rewarded to the extent that it serves the institutional achievement of objectives. In case of conflict, if competing or even conflicting institutional interests (e.g. corporate profit, national security)[40] clash, researchers would have to be prepared to endure negative sanctions (loss of job and status, prison) to comply with this principle. In the sensational study "The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation"[41] 26 leading A.I. developers demand from researchers and developers as a non-delegable task not only to anticipate the consequences of their work but also to actively warn all relevant actors of negative consequences and to constantly expand the circle of those who are informed to participate and make decisions. A dilemma is that, on the one hand, the possible consequences of concrete R&D

[36] As part of the comprehensive European strategy on artificial intelligence, which has been published on 25 April 2018. EU-KOM (2018) and Krempl (2018).
[37] So also: Brundage, et al (2018)
[38] The above-mentioned EU strategy assumes that progress in K.I. development must be monitored by a A.I. observatory. EU-KOM (2018).
[39] Similar: Executive Office of the President (2016), p. 42, Recommendations 19 and 21.
[40] NSTC (2016), p.3.
[41] Ibid.

and applications must be discussed openly at an early stage, formalized and legitimized, and on the other hand, the disclosure of underlying algorithms increases the risk of misuse.[42]

Attempts to develop and apply ethical and legal regulations for A.I. run into nothing, as long as both the dystopian view (taking into account Jonas' concept of the "remote effect of technology": collateral damage as well as effects on the future) is not clear to everyone and a future draft ("how we want to live") has not yet been negotiated. Symptomatic of this is the principle: "Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards". The safety standards (or general legal regulations and limitations) do not yet exist, and it will be difficult to find them, especially as long as technological development is so much faster than legislation can react and provide a framework (the so-called problem of the cultural lag).[43] [44]

3.3 Ethics and values

Principles No. 6 and 7 deal with the requirements of error transparency and operational safety and demand comprehensive safety protection throughout the entire operating time.[45] This raises several initial questions/issues such as:

a) How can the required protection be effectively implemented technically?

b) What should be done in cases of unclear decisions?

c) What if the actions of the A.I. have to be measured by complex, numerous and/or contradictory values whose effective consideration, if at all, would only be possible again with the help of an A.I.?

[42] Brundage, et al (2018).
[43] Ibid.
[44] Despite all the uncertainty, the planned comprehensive European initiative announces the development of a charter for A.I. ethics. This is to be developed in a broad debate from the beginning of 2019. See EU-KOM (2018).
[45] See also: NSTC (2016).

d) A.I. systems will not be clearly locally concentrated, at least in part, but scattered regionally/globally, across several states. In these cases, complete monitoring, security and verifiability are extremely difficult to ensure.

In order to document malfunctions completely and to optimize the chances of success for repairs, design data and source codes must be stored in state-controlled institutions so that they are publicly accessible in the long term. This also includes training data and criteria. Access to source code (and to algorithms) has so far been inadequately regulated. We know that the release of copyrighted source codes can be circumvented relatively easily by economically powerful players. Clear and strict legal regulations and comprehensive research are therefore needed to at least significantly reduce the risk of source code confidentiality. On the other hand, according to current legal situation, the reconstruction of source codes from publicly accessible compiled code (reverse engineering) is generally permissible (although often, as in the EU, limited to certain cases). This in turn facilitates the uncontrolled development of new or improved A.I. outside of known structures with all their possible consequences. Therefore, the demand for the introduction of information rights as well as labelling and publication obligations, such as those imposed by the Federation of German Consumer Protection Organisations (vzbv), is of central importance.[46]

One of the essential characteristics of strong A.I. and further developed weak A.I. is that these show new and unpredictable "behavior", which due to its complexity cannot be reconstructed from the algorithms. The AlphaGo and AlphaGo Zero systems have already demonstrated this exemplarily. The demand for error transparency is therefore contradictory, because some errors will remain inexplicable. This tightens the requirements for control and liability regulations.

The use of A.I. in court proceedings, as mentioned in Principle No. 8, will have to be addressed comprehensively in the future. The motivation for the use already taking place in legal decision-making processes is to achieve efficiency gains, especially in extensive proceedings and to support "better" decisions. A.I. can be used directly in legal decision-making processes to structure the facts of the case, to make a proposal for the decision, to give relapse and social prognoses or in extreme cases to make

[46] Federation of German Consumer Protection Organisations (2017).

an autonomous decision. According to the current legal situation, this is only possible within very narrow limits in most countries.

We see the use of A.I. for evaluation and decision in legal proceedings as problematic.[47] In particular, drawing boundaries in the assumption of causality and defining the limits of free will are a challenge. In the USA, a classification software for relapse and social prognosis is already in use. Initial evaluations show that there is a clear risk that it leads to discrimination based on skin color.[48]

A very probable area of application of the A.I. will be the preparation of expert opinions (e.g. for the credibility of witnesses) or proof to justify the existence or non-existence of a claim. Everywhere in the field of use as evidence a minimum of scientifically founded procedure, transparency and traceability must be observed and proven by the A.I. The methodological means must correspond to the current scientific knowledge of the subject area. Generally speaking, the issue of transparency with regard to the data incorporated and the rules on which it is based is a major problem. In the context of a procedure it is conceivable that already fact finding is directly influenced by realizations, which were won by an A.I.

In individual cases, considerable effort is to be expected in the traceability of the results in order to be able to trace the evidence. Above all, legal regulations are required which, in all cases of objections/appointments/revisions, require a review by higher authorities, which are exclusively staffed by people. In addition, judges (also judicial staff, lawyers, public prosecutors, notaries) must be trained and further educated accordingly and made aware of the dangers of "blind" faith in technology. At this point in time we tend to make a stronger call for a general ban on A.I. in judicial proceedings.

Principle 9 describes responsibilities for advanced A.I. (Advanced AI Systems). It is stated that A.I. designers should assume responsibility for intended and unintended consequences (implications) as "stakeholders". The wording implies that systems that are less advanced are not subject to this

[47] The use of A.I. as a source of information is already controversial, especially with regard to its informative value. For example, the EU Parliament has spoken out in favour of the legal status of robots as an "electronic person". EU Parliament (2017). More than 200 EU scientists and entrepreneurs (especially robotics experts) reacted to this in an open letter: http://www.robotics-openletter.eu/

[48] Angwin/Larson/Mattu/ Kirchner (2016)

responsibility. [49] The question of responsibility (i.e. also causalities) is difficult in complex systems. This problem is exacerbated by globalized systems. In addition, similar cases (e.g. accidents in industrial plants) show that financially strong and powerful interest groups involved are generally committed to preventing the enforcement of liability claims. The multilateral negotiations after 1992, in particular within the framework of the Convention on Biological Diversity (CBD) (e.g. the Biosafety Protocol) have shown that liability issues (in particular the burden of proof) are a decisive and pragmatic regulatory instrument. The question of who is liable for what damage and when, e.g. via risk management strategies and necessary due diligence reviews, has repercussions down to the level of investment decisions.

Principle No. 12 requires the right to personal data protection but restricts it to data generated by users. In contrast, data collected about these users is not included. Thus, by far the largest part of the data remains beyond the control of those to whom this data relates. With regard to A.I., this also includes such personal information that A.I. generates by association/linking/combination of data.[50] The barrier-free and unrestricted right of access to all personal data relating to oneself must be guaranteed. This right must include explicit, prior consent to collection and networking as well as completeness, transparency and the right to cancellation (the right to be forgotten), provided there are no overriding legal reasons for this (e.g. concealment of a crime).[51]

The protection of privacy and personal data is at the heart of Principle No. 13, which states that the use of A.I. may restrict human rights, but not in a manner that is characterized as "unreasonably". A potential restriction of individual and collective freedom must always be assessed in the context of a potential threat to other fundamental values. In principle, data protection and privacy are essential, guaranteed basic values of society, which, especially in the area of A.I., must be designed in such a way

[49] The EP's scientific service, on the other hand, always sees the need to define responsibilities ex ante: EPRS (2016).

[50] Besides the data that people directly or indirectly disclose about themselves (for example through social media, use of search engines or the Internet of Things) and against which everyone could defend themselves, the importance of a collection of data through sensors (starting with surveillance cameras) should not be underestimated. Current developments in the direction of "neuro-data" can also be observed.

[51] The general demand for the right to be forgotten raises legal problems. For example, former employees of the GDR State Security Service tried to have their data erased, which they understandably failed in court.

that they are neither partially nor as a whole restricted without existential hardship. The following foundations are necessary for this:

- Privacy-by-Design or design variants based on this concept, which already take privacy and data protection effectively into account in the technical design process and implement or enable them technically.
- Opt-out and opt-in options should in principle be mandatory. This must also include the linking of data.
- The basic decision on the use of data should be made by the individual. This requires a clear regulation that also makes Big Data transparent and comprehensible for the individual.
- Freedom to Information should be legally designed in such a way that linked data, including the algorithms behind it, must also be provided by the data collectors.
- By means of liability and criminal law, monitoring and data collections by third parties must be designed in such a way that the individual receives effective legal protection against them.

Principle No. 14 wants to "empower" people and reaffirms the goal that A.I. should benefit as many people as possible. "Empowerment" should be understood here as strengthening people's capacity to act. First, it is fundamental that man remains the subject of his actions - with a maximum of self-determination, and his access to economic, social and political participation is not weakened by A.I., but tends to be strengthened. To ensure that as many people as possible benefit from A.I. technologies, appropriate economic, social, political and cultural choices and decisions are needed.

In view of the great promises of A.I., the political demand is legitimate that A.I. makes a visible contribution to reducing economic and social inequality between people. However, the chosen wording of the principle bears the risk that the common good is only understood as the sum of maximized self-interest. Instead, it must be demanded that A.I. must serve the common good. [52]

[52] As part of the planned comprehensive European initiative (EU-KOM (2018), Heise (2018)), a "A.I. on demand" platform is proposed, which, in addition to new A.I. centers of excellence and "digital innovation hubs", should enable small and medium-sized enterprises to participate in development.

Principle No. 15 demands that "economic prosperity created by AI" should be broadly divided. As a political postulate, this ultimately presupposes a different economic system, which, however, is not required in the text and is therefore probably not intended. Despite all caution in the forecasts, initial studies (e.g. by the University of Oxford) show that the fourth wave of production automation made possible by networking and A.I. will lead to the "net" destruction of millions of jobs, even taking into account many newly created jobs.[53] This will primarily affect less qualified people and women for whom no alternatives are available.[54]

Principle No. 18 deals with the danger of an arms race of lethal autonomous weapons systems. These are not generally outlawed. It only admonishes of the danger of an arms race. The USA, the United Kingdom, Russia, China, Israel and South Korea are already developing such weapon systems and they are already stationary on the border with North Korea. In view of the dynamic acceleration in the development of lethal autonomous weapons systems, there are no clear statements on how to prevent an arms race. The use of autonomous units in war or combat situations is a risk, also for the users, due to the lack of ethical behavior and possible missteering.[55] Furthermore, the formulation of Principle 18 shows that military applications of A.I. do not trigger a "prohibition reflex" either.[56]

Leading A.I. researchers also see the danger of aggressive attacks growing rapidly far below the threshold of weapon systems (e.g. through sabotage of vehicles or large-scale installations, as well as the targeted undermining of democracies and state authority), since the costs of such missions are constantly decreasing, and the attacker is extremely difficult to identify.[57] [58]

[53] McKinsey (2017): By 2030, the average gross loss of around 15% (up to a maximum of 30%) of all jobs worldwide (around 400 million jobs; a maximum of 800 million jobs. Frey/Osborne (2013).
[54] Frey/Osborne (2013), p.36ff. WEF (2018).
[55] Scott, et al (2018) und Brundage, et al (2018).
[56] Nevertheless, the Future of Life Institute is committed. During the negotiations on the Convention on Certain Conventional Weapons / Group of Governmental Experts on Lethal Autonomous Weapons Systems in November 2017, they presented a "shock" video that drastically illustrates the dangers of autonomous weapon systems.
[57] Brundage, et al (2018).
[58] The topic is also addressed in: Executive Office of the President (2016), p. 42, Recommendations 22 and 23.

3.4 Long-term problems

Principle No. 19, in the absence of consensus, calls for no assumptions to be made as to whether and what technical limits exist for further A.I. development, thus avoiding the definition of absolute "red lines" for further A.I. development. The precautionary principle, on the other hand, requires exactly that. If we do not know whether and when a strong A.I. will arise, and if we do not know whether and with what consequences this could be dangerous, societies and states must agree on red lines in advance and effectively enforce them.

Principle No. 20 states that advanced A.I. can profoundly change the future of life on planet Earth. Such a fundamental possibility is countered by the demand to plan and develop "appropriately".[59] "Appropriate" is an indefinite legal term - who determines what is "appropriate" and what the standards are? The main problem with the principle is that it is not asked whether it is even permissible, without submission of compelling reasons, to accept profound changes in the history of life on earth, the direction and extent of which people cannot know and foresee. The principle is obviously based on the understanding that everything that is possible will also be done. So, if man is able to develop strong A.I., then he will do so simply because he can.[60]

History, on the other hand, shows - albeit few - exceptions in which societies, states or communities of states have chosen not to pursue certain technology paths because the risks associated with them have been considered unacceptable (e.g. nuclear fission or fossil fuels as energy sources, biological and chemical weapons, CFCs, etc. as coolants). However, this only happened when large parts of the population (especially the elites) with a clear majority refused to (continue to) accept certain dangers associated with the application of these technologies. As a rule, this only happened after undesired consequences occurred. There are also examples that the "technological renunciation" of one actor has not led others to follow the example, but on the contrary to take advantage of it (at present:

[59] The word "history" used here is misleading and does not fit here, since history is backwards. "Course of life" would be more appropriate. The underlying teleological idea of a goal of human existence is at least worth discussing.
[60] See Beck (1986).

peaceful use of nuclear power). However, this cannot justify gross negligence with regard to the precautionary principle.

Principle No. 21 calls for planning and risk mitigation efforts that are appropriate to the expected effects of existential risks and possible disasters. It is important to note that the principle implies that there are existential risks and that disasters are possible. These should be met with the best "efforts". "Efforts" are not the same as success. For example, with the formulation of the principle it is accepted that these efforts, although appropriate in scope to the risk, can ultimately fail.

Contrary to Principle No. 21, the wording of Principle No. 22 is appropriate to the subject. Nevertheless, they also allow R&D and application of "strong" A.I. to continue, provided they are "subject to strict safety and control measures". The possibility of an absolute ban, i.e. a general ban on "strong" A.I., is not considered. Even if such a ban were enforceable, there would still be the danger that somewhere in the world, for the sake of advantage, people would ignore the bans, for example to gain political and/or ideological/religious power. Not only Russian President Vladimir Putin has already publicly stated that the first power to have a "strong" A.I. will rule the world.[61] In this context, particular account must be taken of global networking, with which such systems can gain access to "executive means" such as weapons (immediate threat), facilities usable as weapons, such as nuclear power plants and critical infrastructures, such as power grids (indirect threat).

Principle No. 23 speaks directly of a "super intelligence" that is "to serve the widespread ethical ideals of all humanity and not just one state or one organization". First of all, however, we humans do not live in a worldwide, ideal (paradisiacal) society, which there will probably never be for all life to undergo permanent change. Here the question arises: Why should we want to allow the creation of a super intelligence? Wouldn't we then become believers in a religion that sees man as a necessary intermediate stage of evolution towards higher, digital beings?[62] We cannot and must not want this. A super intelligence could possibly decide the future fate of humanity within the framework of its independently developed basic programming. This would not have to mean that humanity would be

[61] RT.com (2017).
[62] Vgl. Hariri (2017), S. 497ff. und Dotzauer (2017).

destroyed or enslaved, but then we humans would only have "borrowed/granted" opportunities to participate. A violent assertion of human interests against the super intelligence would in all likelihood be doomed to failure.

What these "widespread ideals" are at all remains open. For example, in parts of the world "being rich" or "consuming" or "enjoying" or "having fun" are widespread ideals. But also, literally implementing religious rules is a widespread ideal in parts of the world. Then there are also ideals that are largely excluded worldwide, such as "the women's right of self-determination to have an abortion" and "the protection of the unborn life from conception". We have more than a billion supporters for both. Even the consistent enforcement of the UN Declaration of Human Rights is still rejected by a large number of people, at least in part. So, what is the legitimizing basis of the term "widespread ethical ideals"?

3.5 Conclusion

The authors of the Asilomar Principles and the VDW research team agree that A.I. will profoundly change life on Earth and that the creation of a strong A.I. must be assumed. At least to some extent they also agree that there is a significant potential danger with regard to the further course of human history. The Asilomar principles are an excellent starting point for discussions on how to exploit the potential of artificial intelligence in the coming years. However, the principles are neither an appropriate normative framework for the necessary definition of absolute limits for the research, development and application of A.I. nor for the enforcement of such limits for security purposes.

The VDW research team regards as problematic above all that the principles were developed on the basis of an (implicit) exclusive utopia premise, which assumes that in principle a limitless technology development or application is possible, which only requires regulation in individual cases. Thus, 90% agreement was required in the design of the Asilomar principles to include a regulatory principle. On the other hand, if a risk potential premise were to comply with the EU precautionary principle, 90%

approval would be required for technological development or application. The precautionary principle applicable in the EU should therefore serve as a guide for all further discussions.[63]

4. Recommendations for action

The VDW research team "Technology Assessment of Digitization" is convinced of the need for a broad, targeted social and scientific discussion on how to meet the challenges posed by R&D and artificial intelligence applications. There are initial approaches to this, which in our opinion are still too much dominated by technical enthusiasm and the economic productivity gains that are already visible and will continue to grow significantly in the future. The basically positive view of many voices is often based on the conviction that it can only be a matter of avoiding or at least reducing undesirable consequences of A.I. anyway, but not on whether at least certain applications of weak A.I. and the generation of a strong A.I. should be fundamentally prohibited. Overall, the impression prevails that the promises of A.I. should not be clouded by doubtful or pessimistic observations.

However, this is also due to the fact that the majority of conceivable negative consequences have not yet occurred or are not yet known and, in the case of strong A.I., they will probably be in a future that will last for years or even decades. The dangers are abstract, invisible and largely unknown.

Our imagination about this future is shaped by science fiction literature and films. Warning voices are therefore not yet taken seriously. The future is apparently far away - too far away for political, social debates that revolve around the problems of today and the near future. In the coming chapters we will highlight the most important starting points for what needs to be done.

Scientific Discourse and Research on Technology Assessment of A.I.

[63] If a risk means that catastrophic or existentially threatening events can occur as a result of R&D or the application of A.I. systems and these risks are not negligible, then the potential risks cannot be the subject of "planning and risk mitigation efforts", but must immediately and immediately trigger all necessary measures to avert danger.

Basic research addressing technological consequences and safety issues, especially with regard to human-machine interaction in particular and machine-environment interaction in general, is still only in its infancy. However, there is a considerable need for public research, and research processes must be application-oriented and accompanied by technical development in terms of risk impact assessment, societal interdependencies and future development steps, and must be intensively promoted.

The research funding (from public funds) must launch inter- and transdisciplinary basic research in the fields of law, ethics, social and economic sciences, computer science, but also media sciences, technology and psychology, especially with regard to possible technical designs, both specifically and comprehensively (also for the enforcement of the protection of fundamental rights). [64] [65] The aim must be, among other things, to develop concrete interdisciplinary guidelines for the technical design process, including programming and technology. Cooperations and formalised, transdisciplinary scientific communication processes in research and teaching (e.g. through joint events and publications) will promote rapid penetration, cross-fertilisation and public dissemination of the knowledge gained.

Social and political discourse

Citizens, decision-makers and multipliers must be made familiar with relevant, scientifically based information on A.I. Considerable efforts are needed to process and discuss the necessary information in the relevant fora. In particular, political, legal and economic decision-makers must be enabled to make informed decisions. This also applies to the media and social institutions as well as relevant, subordinate governmental areas (above all the areas of security and consumer protection). Conferences and events should also be used to draw attention to the main questions on the subject of A.I.

[64] Executive Office of the President (2016), p. 42, recommendation 18 also sees the school as having an obligation.
[65] For example, pilot projects should be carried out in a clearly defined space, thus allowing the review of mechanisms of action with limited potentially irreversible consequences, including independent audit mechanisms.

We want to contribute to initiating the necessary discussions and negotiations in order to effectively prevent conceivable dangers of A.I. (especially through strong A.I.) or to minimize them where this is not possible/necessary. It must be taken into account here that a large part of R&D activities on A.I. does not take place under state control; in global competition and, moreover, military-oriented A.I. research is subject to at least limited democratic control. In particular, the involvement of civil society and individuals in guiding and shaping these important issues can therefore be a decisive step towards the success of the necessary measures outlined below.

Regulation

As with all research areas, A.I.'s R&D must follow normative principles (ethical and legal).[66] From the VDW's point of view, the basis must be the Universal Declaration of Human Rights or its codification in national law. Furthermore, existing legal standards such as the precautionary principle must be explicitly extended to technical developments and made legally binding (precautionary principle 2.0).[67] In this sense, risk assessment and technology assessment must be made binding. There is a need for codified rules covering all relevant legal issues at all necessary national and international levels,[68] including prohibitions or moratoria. Standardization must take place within the respective application contexts on the one hand, and in society as a whole on the other. This also includes a call for the precautionary principle to be developed further in legal terms where this is necessary in the interests of comprehensive security. Since the development of internationally valid legal systems equipped with enforcement instruments takes decades rather than years, work must begin immediately.

Effective state (and multilateral) structures and sanction mechanisms are needed to ensure that A.I. is controllable at all times. This applies to all phases of R&D and application. In particular, market launches may only take place after adequate safety assessments have been completed. This requires,

[66] NSTC (2016).

[67] The precautionary principle must be developed further. The precautionary principle and innovation are not mutually exclusive. On the contrary: Precautions for the future in particular should give rise to innovations that promote sustainability. Keeping open and promoting lower-risk alternatives and creating multiple options for the future is far too little. Research and innovation should become aware of their responsibility and values and enter into an early dialogue with society on technology assessment. In the case of risks and opportunities, the omitted alternatives should also be considered.

[68] First considerations in this regard: EPRS (2016).

for example, extensive testing in realistic scenarios. Technical expertise is needed to support this, but neither must it have a decisive influence on regulatory issues nor economic interests. [69] There is also a need for technically informed, comprehensive, globally functioning, democratic control of research and development in this area.

The research funding described above must also address the identification and scientific elaboration of any additional legal models of effective regulation that may be required, which help to identify the effects of A.I. as early as possible and to force all necessary reactions in a timely manner. Committees of experts and representatives of civil society should support this. In addition, ethical self-obligations are required for dealing with A.I. in R&D and application.

The European Commission is launching its paper at the same time as its comprehensive European initiative on A.I., including plans for a European A.I. alliance and an ambitious approach to A.I., which should make the EU a "leader of the A.I. revolution". [70] On the other hand, the strategy's indications of possible distortions sound like forced concessions.[71] There are also initial demands, both by the EU Parliament (recommendations to COM of 27 January 2017) and by the EU Commission itself, to "immediately start all necessary work to develop an internationally recognized legal and ethical framework for the design, production, use and governance of A.I.".[72] This is also specifically demanded by scientists: The EU should immediately develop an A.I. Charter and work towards the development of a global charter.[73]

A.I.-inherent security mechanisms

The creation of any A.I. also requires the irrevocable programming of ethical principles that remain functional in all imaginable modes of operation. In situations where this is not (anyx longer) guaranteed and human intervention is required to prevent damage, all necessary actions must be

[69] Less critical: Executive Office of the President (2016), p. 40, Recommendations 5 and 6.
[70] EU-KOM (2018), p. 14.
[71] Ibid.
[72] EU-DG R+I (2018).
[73] See in detail Metzinger (2018).

possible at all times in a timely and preventive manner. The most important guideline must be that A.I. cannot harm a person under any conceivable circumstances. This complies with the robotic laws of Asimov and is an absolute prerequisite for "usefulness" of the A.I. Asimov himself characterized his laws as necessary but not sufficient.[74] Ethical principles for algorithms are also at least problematic because they do not have an "I-personality" that can have the experience of birth, joy, pain, illness and death. If this were to happen one day, we would face quite different challenges.

Special case of deadly autonomous weapon systems

With regard to the military use of lethal autonomous weapons systems, the German-French (non-)working paper on the first formal UN negotiations on lethal autonomous weapons systems must be built upon.[75] The two EU Member States jointly propose a political declaration at UN level providing for the first steps towards an international protocol under the "Convention on the Prohibition or Restriction of the Use of Certain Conventional Weapons which may cause Excessive Sufferings or have indiscriminate effects".[76] This would in fact be a ban. Here, too, success will depend above all on all states coming to the conclusion that, ultimately, lethal autonomous weapons cannot be safely controlled by their users either. Then only producers of deadly autonomous weapons systems would have an interest in preventing an effective ban. Dr. Alexander Kott, head of the Network Science Division of the Army Research Laboratory, shows how far this path is, calling for massive research efforts in the field of the development of autonomous weapon systems in the USA, since only autonomous weapon systems would be able to react appropriately to the future interaction of the "autonomous Internet of battle things" on the battlefield.[77]

[74] Asimov (1950).
[75] The negotiations took place in November 2017 as part of the Convention on Certain Conventional Weapons: Group of Governmental Experts (GGE) on lethal autonomous weapons systems (LAWS). Group of Governmental Experts of the High Contracting Parties (2017) und International Committee of the Red Cross (2004). For the position of the P.R. China see: Kania (2018).
[76] United Nations (1980).
[77] Kott (2018).

Invitation to dialogue

The tasks ahead of us can only be mastered together. We therefore offer our support to all those who are prepared to use the opportunities and possibilities of A.I. only to the extent that they do not endanger human health, life or the environment and do not harm the common good.

Below the threshold of existential risks, there will be, and justifiably so, an intensive social and political debate about what benefits the common good, or at least what does not harm it or is to be regarded as harmful. We look forward to entering into a broad dialogue on this issue.

5. Literature

Angwin, Julia/ Larson, Jeff/ Mattu, Surya/ Kirchner, Lauren (2016): Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks, in ProPublica, May 23, 2016; https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Asimov, Isaac (1950): I, Robot. Gnome Press.

Bahr, Egon/ Lutz, Dieter S. (Hrsg.) (1992): Gemeinsame Sicherheit. Idee und Konzept. Bd. I: Zu den Ausgangsüberlegungen, Grundlagen und Strukturmerkmale Gemeinsamer Sicherheit, Baden-Baden.

Beck, Ulrich (1986): Risikogesellschaft. Auf dem Weg in eine andere Moderne. Erstausgabe, 1. Auflage. Suhrkamp, Frankfurt am Main.

BITKOM (2017): Künstliche Intelligenz verstehen als Automation des Entscheidens – Leitfaden, Berlin.

Brundage, Miles, et al (2018): The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation; https://www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf.

Centre on Impact of AI and Robotics (der UNSW Sydney) (2018): Open Letter to Professor Sung-Chul Shin, president of KAIST from some leading AI researchers in 30 different countries; https://www.cse.unsw.edu.au/~tw/ciair/kaist.html.

Dotzauer, Gregor (2017): Näher, mein Bot, zu dir. In: Tagesspiegel 12.12.2017.

Eberle, Ute (2015): Sprachassistenten verändern unser Leben. In: Wirtschaftswoche 28.7.15.

EFI (Expertenkommission Forschung und Innovation) (2018): Gutachten zu Forschung, Innovation und technologischer Leistungsfähigkeit 2018, Berlin.

EPRS (2016): European Parliamentary Research Service, Scientific Foresight Unit (STOA), PE 563.501: Ethical Aspects of Cyber-Physical Systems, Scientific Foresight study, Brüssel.

EU-DG R+I (2018): European Commission, Directorate-General for Research and Innovation; European Group on ethics in Science and New Technologies: Statement on Artificial Intelligence, Robotics and "Autonomous" Systems, Brüssel.

EU-KOM (2018): Maximising the benefits of Artificial Intelligence (Version 15 – 27/02/2018). Unveröffentlichtes Arbeitsdokument, Brüssel.

EU-Parlament (2017): European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics; http://www.europarl.europa.eu/sides/getDoc.do?type=TA&reference=P8-TA-2017-0051&language=EN&ring=A8-2017-0005.

Executive Office of the President (2016): Preparing for the Future of Artificial Intelligence. Washington, D.C.

Experts on Lethal Autonomous Weapons Systems (LAWS), Geneva, CCW/GGE.1/2017/WP.4; http://undocs.org/ccw/gge.1/2017/WP.4.

Frey, Carl Benedikt/ Osborne, Michael A. (2013): The Future of Employment: How Susceptible are Jobs to Computerisation, Oxford University.

Future of Life Institute (2017): Asilomar AI Principles; https://futureoflife.org/ai-principles/ und Beneficial AI 2017. Conference Schedule; https://futureoflife.org/bai-2017/.

Göttinger Achtzehn (1957): Göttinger Manifest; https://www.uni-goettingen.de/de/text+des+g%c3%b6ttinger+manifests/54320.html.

Group of Governmental Experts of the High Contracting Parties (2017): To the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects. For consideration by the Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS). Submitted by France and Germany, CCW/GGE.1/2017/WP.4, Geneva.

Harari, Yuval Noah (2017): Homo Deus – Eine Geschichte von Morgen, München.

Henk, Malte (2014): Jugend ohne Sex. In: Zeit online vom 15.6.14.

Independent Commission on Disarmament and Security Issues (1982): Common security: A blueprint for survival, Simon and Schuster, New York.

International Committee of the Red Cross (2004): Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Geneva.

Jonas, Hans (1979): Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation. 1. Auflage, Insel-Verlag. Frankfurt am Main.

Kania, Else (2018): Artificial Intelligence: China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems. In: Lawfare, April 17, 2018; https://www.

lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems.

Kommission der Europäischen Gemeinschaften (2000): Mitteilung der Kommission: die Anwendbarkeit des Vorsorgeprinzips, KOM (2000) 1 endgültig, Brüssel; https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52000DC0001&from=DE.

Kott, Alexander (2018): Challenges and Characteristics of Intelligent autonomy for Internet of Battle Things in Highly Adversarial environments, Adelphi, MD; https://arxiv.org/ftp/arxiv/papers/1803/1803.11256.pdf.

Krempl, Stefan (2018): Künstliche Intelligenz: EU-Kommission plant umfassende europäische Initiative. In: heise online 26.3.2018.

McKinsey (McKinsey Global Institute), (2017): Jobs Lost, Jobs Gained: Workforce Transitions in a time of Automation.

Metzinger, Thomas (2018): Towards a Global Artificial Intelligence Charter. In: European Parliamentary Research Service: Should we fear artificial intelligence? Brüssel.

Müller, Vincent/ Bostrom, Nick (2013): Future progress in artificial intelligence: A Survey of Expert Opinion. In: Vincent C. Müller (Hrsg): Fundamental Issues of Artificial Intelligence, Berlin.

NSTC (National Science and Technology Council) (2016): the National Artificial Intelligence Research and Development Plan, Washington, D.C.

Oye, Kenneth A., et.al. (2014): Regulating gene drives. Regulatory gaps must be filled before gene drives could be used in the wild, in: Science 17 Jul 2014; http://science.sciencemag.org/content/early/2014/07/16/science.1254287.full.

RT.com (2017): 'Whoever leads in AI will rule the world': Putin to Russian children on Knowledge Day, 1 Sep, 2017 14:08; https://www.rt.com/news/401731-ai-rule-world-putin/.

Schellnhuber, Hans Joachim (2015): Selbstverbrennung: Die fatale Dreiecksbeziehung zwischen Klima, Mensch und Kohlenstoff, C. Bertelsmann Verlag, München.

Scott, Ben/ Heumann, Stefan/Lorenz, Philippe (2018): Artificial Intelligence and Foreign Policy. Stiftung Neue Verantwortung, Berlin.

TAB (Büro für Technikfolgenabschätzung beim Deutschen Bundestag) (2016): Technologien und Visionen der Mensch-Maschine-Entgrenzung. Sachstandbericht zum TA-Projekt „Mensch-Maschine-Entgrenzungen: zwischen künstlicher Intelligenz und Human Enhancements. Arbeitsbericht Nr. 167, Berlin.

United Nations (1948): Universal Declaration of Human Rights, Paris; http://www.un.org/en/universal-declaration-human-rights/.

United Nations (1966): International Covenant on Economic, Social and Cultural Rights. Adopted and opened for signature, ratification and accession by General Assembly resolution 2200A (XXI) of 16 December 1966. Entry into force 3 January 1976 in accordance with article 27; https://www.institut-fuer-menschenrechte.de/fileadmin/user_upload/PDF-Dateien/Pakte_Konventionen/ICESCR/icescr_en.pdf.

United Nations (1980): Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be deemed to be Excessively Injurious or to have Indiscriminate Effects (with Protocols I, II and III), Geneva, 10 October 1980; http://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVI-2&chapter=26&lang=en.

United Nations (1992): Agenda 21. Konferenz der Vereinten Nationen für Umwelt und Entwicklung, Rio de Janeiro; http://www.un.org/Depts/german/conf/agenda21/agenda_21.pdf.

United Nations (2015a): Paris Agreement, Paris; https://unfccc.int/files/meetings/paris_nov_2015/application/pdf/paris_agreement_english_.pdf.

United Nations (2015b): Transforming our world: the 2030 Agenda for Sustainable Development. Resolution adopted by the General Assembly on 25 September 2015. A/RES/70/1; http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E.

Université de Montréal (2018): The Declaration: https://www.montrealdeclaration-responsibleai.com/the-declaration.

Verbraucherzentrale Bundesverband (2017): Algorithmenbasierte Entscheidungsprozesse - Thesenpapier des vzbv, Berlin.

Vowinkel, Bernd (2017): Ist der Mensch eine Maschine; https://transhumanismus.wordpress.com/2017/06/14/ist-der-mensch-eine-maschine/.

Walter, Sven (2014): Situierte Kognition. In: Information Philosophie; Heft 2/2014, S. 28-32, Lörrach.

Weizsäcker, Ernst Ulrich von/ Wijkman, Anders (2017): Wir sind dran. Club of Rome: Der große Bericht. Was wir ändern müssen, wenn wir bleiben wollen. Eine neue Aufklärung für eine volle Welt, Gütersloh.

WEF (World Economic Forum in collaboration with The Boston Consulting Group), (2018): Towards a Reskilling Revolution: A Future of Jobs for All, Köln/Genf.

6. Annex

6.1 Asilomar AI Principles

Research Issues

1) Research Goal: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.

2) Research Funding: Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as:

- How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
- How can we grow our prosperity through automation while maintaining people's resources and purpose?
- How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?
- What set of values should AI be aligned with, and what legal and ethical status should it have?

3) Science-Policy Link: There should be constructive and healthy exchange between AI researchers and policy-makers.

4) Research Culture: A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.

5) Race Avoidance: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

Ethics and Values

6) Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

7) Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.

8) Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

9) Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

10) Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

11) Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

12) Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.

13) Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

14) Shared Benefit: AI technologies should benefit and empower as many people as possible.

15) Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

16) Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

17) Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.

18) AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

Longer-term Issues

19) Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.

20) Importance: Advanced AI could represent a profound change in the history of life on Earth and should be planned for and managed with commensurate care and resources.

21) Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.

22) Recursive Self-Improvement: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.

23) Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.

6.2 Authors (Members of the VDW Research Team)

Prof. Dr. Ulrich Bartosch (political scientist) is professor of education at the Catholic University Eichstätt-Ingolstadt, where he teaches and conducts research on pedagogical theory and the history of political ideas, university reform and education, competence description and development, inclusion, participation, school social work and world internal politics. From 2005-2011 he was chairman of the Social Work Day. He is a member of the Ad-hoc-WG "Crediting and recognition of digital learning formats" of the Hochschulforum Digitalisierung der Hochschulrektorenkonferenz (HRK) and head of the cooperation project of the Catholic University of Eichstätt-Ingolstadt (KU) and VDW "Laudato Si' - Die päpstliche Enzyklika im Diskurs der Große Transformation". From 2009 to 2015 he was Chairman of the Association of German Scientists and has been Chairman of the Advisory Board of VDW ever since.

Prof. Dr. Stefan Bauberger SJ (physicist, philosopher) is Professor of Natural Philosophy and Philosophy of Science at the University of Philosophy in Munich. He is a member of the Jesuit order and theologian. He holds a doctorate in theoretical physics and a habilitation in philosophy. He worked for several years in theoretical elementary particle physics. He researches and teaches about border issues between philosophy and natural science, especially physics, in the field of the dialogue between science and religion as well as the philosophy of Buddhism, and in the fields of philosophy of technology and science theory. Before that he was head of the formation of the Jesuit order in Germany. He is a ZEN Master and runs a meditation center.

Tile von Damm (political scientist) is head of the MOD urban research institute and associate researcher at the TU Berlin. He is an Urban Expert in the EU project "Orfeo&Majnun". He is also co-founder and managing director of DiMed to ensure rural basic medical care. His research and work focuses on inclusive rural and urban development, participation and global governance, open source and open data, and research development and transfer. He is a member of the European network on culture and creative industries. He was research manager at the Centre for Literary and Cultural Research (ZfL), head of the PerGlobal research institute and coordinator of the Excellence Initiative at Humboldt University Berlin. At the UN World Summit on Sustainable Development and in the UN negotiations on the information society, he was part of a civil society negotiating delegation.

Dr. Rainer Engels (agronomist) is an expert on economic policy in development cooperation and has been working on questions of development economics for many years. His focus is on the developmental design of trade, investment and industrial policy, in particular intellectual property rights and technical standards. Since 2015 he has been working on the automation and digitization of industrial production (Industry 4.0) and electromobility. He is a consultant for sustainable economic policy and private sector promotion at GIZ. Previously, he was Managing Director of Germanwatch for many years.

Prof. Dr. Malte Rehbein (historian) holds the chair for Digital Humanities at the University of Passau, where he researches and teaches formal and computer-based methods including A.I.-based methods and their possible applications for tasks and questions in the humanities and cultural studies with a

special focus on history. He publishes on Historical Data Studies, Digitisation of Cultural Property and Data Modelling as well as on questions of ethics and criticism of science and society. Professional experience includes the IT and consulting industry; academic stations were Göttingen, Würzburg, Galway/Ireland, Victoria/Canada and Lincoln-NE/USA. He is a member of the Historical Commission of the Bavarian Academy of Sciences. Since 2017 he has been a member of the Association of German Scientists.

Frank Schmiedchen (Economist, MBA) is Government Director at the Federal Ministry for Economic Cooperation and Development (BMZ); responsible, among other things, for the IMF and international economic and financial issues. Previously, he was responsible for biodiversity and biosafety, foreign and security policy issues in Africa, ACP, Industrial Policy, UNIDO, Intellectual Property Rights and the development of local pharmaceutical production at BMZ and the Permanent Representation of the Federal Republic of Germany to the EU. Previously, he was Dean of the Department of SME Management at the Pontifical Catholic University of Ecuador and coordinated the relevant departments for the Association of Jesuit Universities in Latin America (AUSJAL). 2002-2009 and since 2016 he was/is a member of the Advisory Board of the Federation of German Scientists. Since October 2017, he has headed the VDW Technology Assessment of Digitisation Research Team.

Prof. Dr. Heinz Stapf-Finé (sociologist, economist) is Professor of Social Policy at the Alice Salomon University Berlin and Academic Director of the Paritätische Akademie Berlin. He holds a doctorate on the subject of "old-age provision in Spain" and is an international expert in the field of labour and social policy. Before his appointment as a university lecturer, he was Head of the Social Policy Division of the Federal Executive Committee of the Federation of German Trade Unions (DGB). He gained his first professional experience as Operations Manager of the Luxembourg Income Study and as a research assistant at the Institute for Health and Social Research (IGES) Berlin. He then worked as a policy officer for the German Hospital Federation. He has been a member of the Association of German Scientists since 2006.

Angelika Sülzen (MBA) is Government Director at the Federal Ministry for Economic Cooperation and Development (BMZ), where she is currently responsible for the areas of equality, balance of work and family life and health management. Previously, she was responsible for bilateral cooperation between the Federal Republic of Germany and Burundi respectively the Central African Republic. Prior to that, she was responsible for budget and financial issues and led a large IT project at BMZ. From 2003 to 2007 she worked for the German Development Service in South Africa and Lesotho.

We thank them for their contributions and suggestions: Lucas Bartosch, Judith Buttenmüller, Prof. Dr. Hartmut Graßl, Prof. Dr. Regine Kollek, Dr. Hans-Jochen Luhmann, Dr. Michael Marhöfer and Christine von Weizsäcker. We thank for quality control of the English version: Georgina Ellis.